14 ALTERNATIVES AND MODIFICATIONS

14.1 Alternative Statistical Tests

The nonparametric statistical tests described in this report are expected to perform well in a wide variety of circumstances. However, in some situations alternative methods can be considered. As mentioned in Section 2.4.1, there are many statistical tests that can be used for determining whether or not a survey unit meets the release criteria. Any one test may perform better or worse than others, depending on the hypotheses to be tested, i.e., the decision that is to be made and the alternative, and how well the assumptions of the test fit the situation. Some possible alternatives are discussed below.

In evaluating statistical tests, generally one chooses the test that has the highest power among the various alternatives. The power is compared when each test is set to have the same Type I error rate, α. The Type I error rate is the probability that the null hypothesis will be rejected when it is true. If the assumptions made about the data distribution are correct, the calculation of α forms the basis for setting the critical value of the test statistic. If the assumptions are not valid, the calculated value of α will differ from the true Type I error rate. The fewer assumptions that are made, the more confidence can be placed in the calculation of the Type I error rate.

If a specific set of assumptions is made, the test results can be simulated using Monte Carlo sampling techniques. Using a large number of simulations, the actual Type I and Type II error rates for different tests can be compared. For each sample size, and specific set of assumptions, a separate simulation must be performed. Although much can be learned about the relative accuracy of statistical tests in this way, it is clearly not possible to explore every potential set of assumptions.

An alternative is to look at large sample results. With very few exceptions, it can be proved that the average of a large enough number of random data points tends to be normally distributed (Central Limit Theorem). In the same way, the power of statistical tests can be examined when the sample sizes are large enough. Note that what is meant by large enough is not precisely specified. Depending on the situation, large enough might be 10, or it might be 1000. If the samples size is allowed to grow large enough, the asymptotic (i.e., in the limit of arbitrarily large sample size) behavior of tests can be compared. Better large sample test behavior may be taken to imply that a test is better for all sample sizes. In reality, it can only be used as an indication of which test might be preferred.

One measure commonly used to compare statistical tests is called the relative efficiency. This is defined as the inverse of the ratio of sample sizes needed to achieve a given level of statistical power. If a test has relative efficiency of two relative to another, it requires half the sample size to achieve the same power. The asymptotic relative efficiency of one test to another, is the limit of the relative efficiency when the sample size is arbitrarily large.

Wilcoxon Signed Ranks Test (WSR test)

The asymptotic relative efficiency of the WSR test compared to the Sign test can be greater or less than one. That is, either might be better, depending on the data distribution. The WSR test tends to be better when the data distribution is symmetric, and the Sign test tends to be better when the data distribution is skewed.

Student's t-Test

Student's t-test may be used if the data have a normal distribution. This is a more restrictive requirement than that of symmetry, since every normal distribution is symmetric, but there are many other distributions that are also symmetric. The assumption of normality should be checked before using this test. The Shapiro-Wilk test discussed in EPA/QA-G9 (1996) is one such test. Others include the Kolmogorov-Smirnov test, Lillifor's test, and the Chi-Squared test.

The asymptotic relative efficiency of the WSR test relative to the one-sample Student's t-test ranges from 0.864 to infinity. As stated by Conover (1980): "the Wilcoxon test never can be too bad, but it can be infinitely good... ." The asymptotic relative efficiency of the WRS test compared to the two-sample Student's t-test has the same range, from 0.864 to infinity.

Chen's Test

Chen's test (Chen, 1995) is a modification of the Student's t-test that has been suggested for use when data are from a positively skewed distribution. Simulations show that it is generally more powerful than other forms of the t-test. However, this test can only be used in Scenario B.

Lognormal Test

If the data are assumed to lognormal, the testing procedure of Land (1988) may be used. The assumption of lognormality should be checked by testing the logarithms of the data for normality. It is important to note that a test on the mean of a lognormal distribution *cannot* be performed by using a Student's t-test on the mean of the logarithms of the data. This is because the mean of the logarithms of the data is the logarithm of the *median* of the original data. The behavior of this test relative to others when the assumption of lognormality is violated has not been studied.

Bootstrap Methods

The bootstrap is a simulation technique (Efron and Tibshirani, 1993). In essence, the distribution of concentrations in a survey unit is approximated by the empirical distribution (e.g., histogram) of the sample data taken. If n measurements are made, these n measurements are randomly sampled n times with replacement. Each time this is done, the mean of the random sample is calculated. After this has been done a specified number of times (generally between 50 and 200), the standard deviation of all of the random sample means is calculated. This is then used as the estimate of the standard error of the mean. There are, in addition, several methods for computing bootstrap t-statistics. Usually 1000 or more replications are recommended for the bootstrap t. Bootstrap methods generally have good asymptotic properties, but can be sensitive to outliers, and erratic when sample sizes are small.

14.2 Retesting

It may happen that a survey unit fails the hypothesis test (i.e., the decision is made that the survey unit does not meet the release criterion), yet the mean of the measured data is below the release criterion. This is more likely to occur when the mean falls in the gray region than otherwise. It analogous to the situation in which the mean is below the release criterion, but the $1-\alpha$ upper confidence level on the mean falls above the release criterion. It may be that the survey unit does meet the release criterion, but the hypothesis test was not powerful enough to detect that with the number of samples taken. Under some circumstances, one might like the option to take additional random samples and re-perform the hypothesis test on the entire set of data. The major difficulty with this is that the Type I error rate will now be greater than originally specified in the DQOs.

Sequential testing is performed when data are collected and analyzed in stages. It differs from hypothesis testing in that at each stage a third alternative is added to the decision of whether or not to reject the null hypothesis, namely, to collect more data before deciding. The usual motivation for sequential testing is to reduce the expected total number of samples from that required when all the sample are taken at one stage. Sequential versions of the WSR and WRS tests are discussed by Spurrier and Hewett (1976).

14.3 Composite Sampling

The number of measurements taken in a Class 1 survey unit may sometimes be driven more by the need to locate small areas of elevated activity than by the need to achieve the specified acceptable error rates for the statistical tests. When the scanning MDC is high, the sample size, N, may need to be significantly increased, in order to decrease the area between samples on the systematic grid. When this grid area, about A/N, is small enough, so that in turn the area factor is sufficiently high, the result is a $DCGL_{EMC}$ that is detectable by scanning. If the sample size N is much greater than that required for the statistical tests, some number, m, of neighboring samples might be composited to reduce the total cost of analysis. Suppose there are $N = mn$ measurements. Each composite represents a contiguous area of approximately the same proportion of the survey unit, $m(A/N)$. The number of composite measurements, n, should be equal to or greater than the number of measurements required by the statistical test. When the elevated measurement comparison is performed against the composites measurement results, the $DCGL_{EMC}$ should be divided by the number of samples included in each composite. If the composite measurement is below $DCGL_{EMC}/m$, no individual sample contributing to the composite could exceed the $DCGL_{EMC}$.

If a composite measurement is flagged by the EMC, it may be necessary to reanalyze each sample included in that composite to determine which of them, if any, actually exceed the $DCGL_{EMC}$ or, alternatively, the area of the survey unit represented by that composite measurement should be reinvestigated.